THUIR at TREC2008: Enterprise Track[1]

Yufei Xue, Tong Zhu, Guichun Hua, Min Zhang, Yiqun Liu, Shaoping Ma
State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China
Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China
z-m@tsinghua.edu.cn

Abstract. We participate in document search and expert search of Enterprise Track in TREC2008. The corpus and tasks are same as the year before. Different from TREC 2007, the topics come from CSIRO Enquiries, and the topic statements are richer and more colloquial.. In document search, we look into the key resource page pre-selection, the use of anchor text, query classification, and multi-field search. In expert search, we develop methods to detect expert identifiers and experimented based on our previous PDD (personal description documents) model.

1 Introduction

This is the fourth year that the IR groups of Tsinghua University participated in TREC Enterprise Track. Different from TREC 2007, the topics come from CSIRO Enquiries, and the topic statements are richer and more colloquial. The approaches we've studied this year include the use of anchor text, person entity identification, topic distillation with key resource pre-selection, query classification and multi-field search.

For document search task, we mainly investigate the effects of key source pre-selection and the use of anchor text. We first observe the high quality resource distribution. Some features are studied to find overview pages. We also do some link analysis: both HITS and PageRank algorithms are employed to evaluate the page quality. Besides, we attempted a novel link analysis method which involved the document similarity.

For expert finding task, a lot of efforts have been made on name identification. We built personal description documents (PDD) for each candidate from various types of resources. We obtain retrieved results from each description document collection.

2 Document Search

The document search task is to help the science communicator to find specific information in the web site. We build a query independent classifier that selects key resources to find high quality pages. The other one is to adapt link analysis to predict those authoritative pages. Both approaches were adopted in our experiments.

2.1 Key pages pre-selection using query independent features

First we try to do some data cleansing work to pick those key pages out according to some query independent features. Figure 1 shows that the distribution of the amount of out-links in overview pages provided by TREC. From the distribution we notice that most overview pages contain a lot of out-links from 115 to 180, contrast to the fact that among the whole corpus more than ninety percent of pages are with out-links less than 100. We conduct an experiment that only retrieve from those documents which have number of out-links more than 100 and build a training set which is not overlapped with the Enterprise 2008 topics. The results show that the performance is 10% higher than retrieving from the whole corpus while the size of the selected corpus is only 10% of the whole one.

[1] Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108)，Natural Science Foundation (60621062, 60503064，60736044) and National 863 High Technology Project (2006AA01Z141)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
THUIR at TREC2008: Enterprise Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Tsinghua University,State Key Laboratory of Intelligent Technology and Systems,,Beijing 100084, China,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**4**	

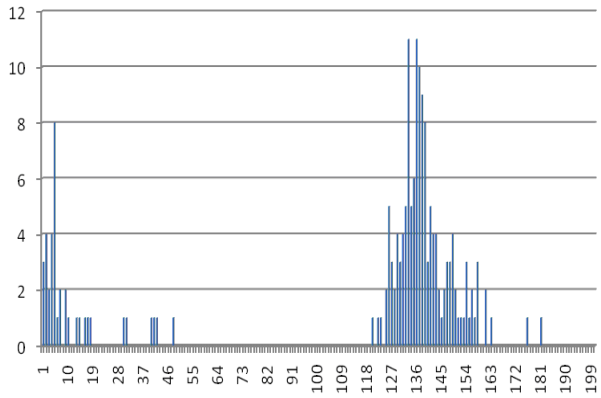

Figure 1. The distribution of the amount of out-links in example pages

Similarly, the in-link anchor text are also studied. We merge the distillation of the page A and all its anchor texts whose links leads to the page A. The new page B replaces the page A in the whole process of the information retrieval. The experiments results in our training set validate our supposition that the anchor text based retrieval outperform the full text retrieval by more than 20% in MAP.

Some other features are also attempted to identify high quality pages, such the length of page URLs, the amount of in-link. We also implement a decision tree to combine these features to better construct the high quality pages set.

2.2 Adapting link analysis for finding authoritative pages

We first used Page Rank [1] to estimate the quality of the web pages. However, the link structure of the web site CSIRO is quite different from the environment that the algorithm applies to. So after executing the Page Rank process, we get some pages with very large scores like the homepage and index pages. However, it is difficult to use these importance scores in conjunction with query-specific IR scores to rank the query results. Several known approaches are attempted but all failed to improve the performance.

Assume that if one page provides necessary authoritative information to one topic, then its linked neighbors will have higher probabilities to present some detailed or relevant information. Therefore for a given page, we add the similarities of those pages which the link to the page as the new score. This reflects the idea that considering the similarity of pages instead of merely analyzing the link relation. For example, if a hub points to an authoritative page and the hub page is a good one, then the authoritative may get a lot of reinforcement. However, if only a little part of the hub page is relevant to the page it links to, the authoritative page may get too much. Plus the strength of the link relation is more or less reflected by the anchor text. Therefore involving the similarity of anchor text to the algorithm may better quantitatively determine the strength of reinforcement. The experiments results show that the improvement achieved by the content-based link analysis is consistent and significant.

2.3 Submitted results

We list the four results we submit and their descriptions in the following table.

	Brief Description
THUFmfS	short query, result combination on fulltext and selected key pages set.
THUFS	short query search on full text.
THUFaAS	short query search on full text + anchor text + selected key pages.
THUFsimAncL	long query, result reranking with inlink similarity, then reranking with anchor evidence

We adapted short queries and rank the combination with probabilistic relevance model on original full text and the pre-selected key pages set on THUFmfS. The result of THUFS is the baseline result. We extract and

integrate inlink anchor text for each document, combine them into original documents, and rank the documents with probabilistic relevance model on the new collection in THUFaAS. In the last result of THUFsimAncL, we use the probabilistic relevance model to rank the linear combination of the relevance between the long queries and the documents, and the similarity propagation with in-link docs. After that we re-rank the result with the anchor text similarity evidence.

3 Expert Search

3.1 Expert identifiers detection

Because there is no master list of candidates, we should automatically detect expert identifiers. According to the guideline that among CSIRO the pattern of the emails is firstname.lastname@csiro.au, we extract all the email addresses from the corpus, including some variation such as that the @ symbol may be HTML-encoded. After eliminating the typo in name and name variations in addresses, we got 3161 candidate.

There are some variations for English names. Given a name “firstname.lastname“, at least five variations are possible: Firstname Lastname, Firstname.Lastname, Firstname Middlename Lastname, F. Lastname, F. M. Lastname. For the first three pattern of variation, we use Aho-Corasick algorithm to label. It takes O(m+n) time, where m is the length of the name and n is the length of the document. Labeling F.Lastname is a problem that a bit hard to tackle. Because it is possible that more than one person shares one abbreviation. For example, T. Thomas may represent Tom Thomas or Tim Thomas. We tried to eliminate the ambiguity according the co-occurrence of other labeled names. However, there are still some ambiguous abbreviations hard to eliminate.

Some other technologies such as pronouns eliminating are also integrated in our system.

3.2 Constructing PDD and merging results

As we did in previous expert finding task, we build person description document (PDD) for each candidate [4]. We extract some candidate relevant information from the document as expertise, for example, the context around the expert identifier.

Besides, we also extract information from candidate’s homepages. In total we find 477 homepages, which are about 15.2% of the amount of candidates. We name the PDD constructed from the homepages as detailed PDD (DPDD). We also construct another PDD using a collection of high-quality documents. The documents in the collection are selected by analyzing the link graph of the corpus.

For job hunting requests, we filtered out all the documents in the domain name of recruitment, and build a recruitment-PDD to improve the performance on the topics about job hunting. And the corresponding automatic query classification approach is used to identify such tasks. But due to the free expressions on find a job in human language, the recall of query classification result is not very high.

To merge the ranking results retrieved from each PDD collection, we adopted EM algorithm to assign the weight to each ranking similarity, because some previous work shows that when the ratio of the weights is parallel to the ratio of the MAP achieved by each ranking list, the merged list achieves the best performance [5].

To compare the performances of two branches of work: PDD-based search (first extract information, then search on re-constructed documents) and document-based search (first search on original documents, then extract expert information from top-returned results), besides runs with PDDs, one run of experiments on document-based search was also submitted.

3.3 Submitted results

The results of our 4 runs are listed below.

Run No.	Run Tag	Description
1	THUPDDSwp	Baseline result. Document refinement to construct person description documents (PDD). Experts search on PDD with word-pair-based probabilistic relevance ranking model.
2	THUPDDSlL	Experts retrieved from the result of document search task, using link analysis to improve. Narrative fields were used. Combined with 1
3	THUPDDlcS	Combination of 3 results: (1) Baseline result (run 1); (2) run 2; (3) search on PDD built with selected high quality pages.
4	THUPDDlchrS	Run 3 + Automatic query type classification.

4 Discussion and Future Work

In document search task, we attempted a promising content-based link analysis algorithm. We find the link structure of intranet and a website is totally different from the Internet. Therefore we will try to investigate the link analysis algorithms applying in intranet. A bigger ambition is that we would like to experiment our content-based link analysis algorithm on bigger data sets and to propose an algorithm integrating the link analysis and similarity ranking.

The automatic query type identification is also important in Intranet search to meet different information request. Further study will be made on the analysis the intention of the query.

References

1. L. Page, S. Brin, R. Motwani, and T. Winograd (1998). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, Stanford, CA.
2. J. M. Kleinberg (1998). Authoritative sources in a hyperlinked environment. Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms.
3. A.P.Dempster, N.M. Laird, and D.B. Rubin. Maximum-likelihood from incomplete data via the em algorithm. J. Royal Statist. Soc. Ser. B., 39, 1977.
4. Fu Yupeng et al “THUIR at TREC 2005: Enterprise Track” State Key Lab of Intelligent Tech. & Sys., CST Dept, Tsinghua University, Proceedings of TREC2005, NIST, 2005
5. Dayong Ding and Bo Zhang. Probabilistic Model Supported Rank Aggregation for the Semantic Concept Detection in Video. ACM International Conference on Image and Video Retrieval (CIVR 2007), July 9-11 2007